(19) JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: **11085193 A**

(43) Date of publication of application: **30.03.99**

(51) Int. Cl **G10L 5/04**
G10L 3/00
G10L 9/18

(21) Application number: **09248750**

(22) Date of filing: **12.09.97**

(71) Applicant: **SANYO ELECTRIC CO LTD**

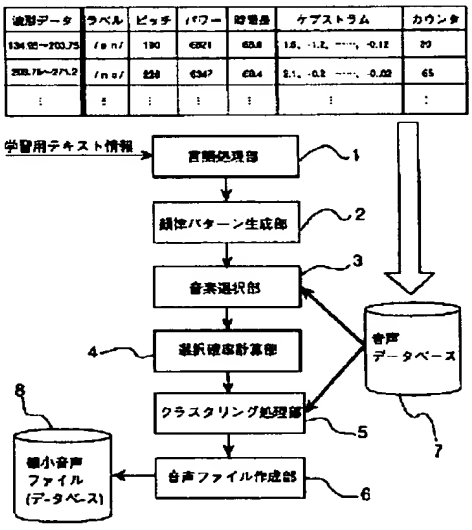
(72) Inventor: **HIRAI HIROYUKI**
ONISHI HIROKI
NISHIDA HIDEJI
HASHIMOTO MAKOTO

(54) **PHONEME INFORMATION OPTIMIZATION METHOD IN SPEECH DATA BASE AND PHONEME INFORMATION OPTIMIZATION APPARATUS THEREFOR**

(57) Abstract:

PROBLEM TO BE SOLVED: To form a speech file from phoneme information of high use frequencies by determining the use frequencies of each pieces of phoneme information constituting a speech data base and executing clustering processing in accordance with these use frequencies.

SOLUTION: A rhythm pattern forming section 2 estimates the basic frequency F_0 near the phoneme center, power and phoneme duration time by using the part-of- speech information of the input text obtainable from the results of phoneme symbol, accent symbol string and morphological analyses. Next, a cost is determined and the phonemes are selected. The text information (sentence) for learning is synthesized and the number of the selected times of the respective phonemes is calculated by using the speech data base 7 including all the phonemes. Next, the number of the selected times of the phonemes included in the respective phoneme units is averaged to form the number of selected times of the phoneme units and the selection probability of the respective phoneme units is calculated. A clustering processing section 5 executes the clustering processing to the number assigned by using the LBG algorithm in the case the distribution is known in accordance with the calculated selection probability.



(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平11－85193

(43)公開日　平成11年(1999)３月30日

(51)Int.Cl.[6]	識別記号	ＦＩ	
G10L	5/04	G10L	5/04 E
	3/00		3/00 H
	9/18		9/18 E

審査請求　未請求　請求項の数９　ＯＬ　（全８頁）

(21)出願番号　特願平9－248750

(22)出願日　平成９年(1997)９月12日

(71)出願人　000001889
三洋電機株式会社
大阪府守口市京阪本通２丁目５番５号
(72)発明者　平井　啓之
大阪府守口市京阪本通２丁目５番５号　三洋電機株式会社内
(72)発明者　大西　宏樹
大阪府守口市京阪本通２丁目５番５号　三洋電機株式会社内
(72)発明者　西田　秀治
大阪府守口市京阪本通２丁目５番５号　三洋電機株式会社内
(74)代理人　弁理士　安富　耕二　（外１名）

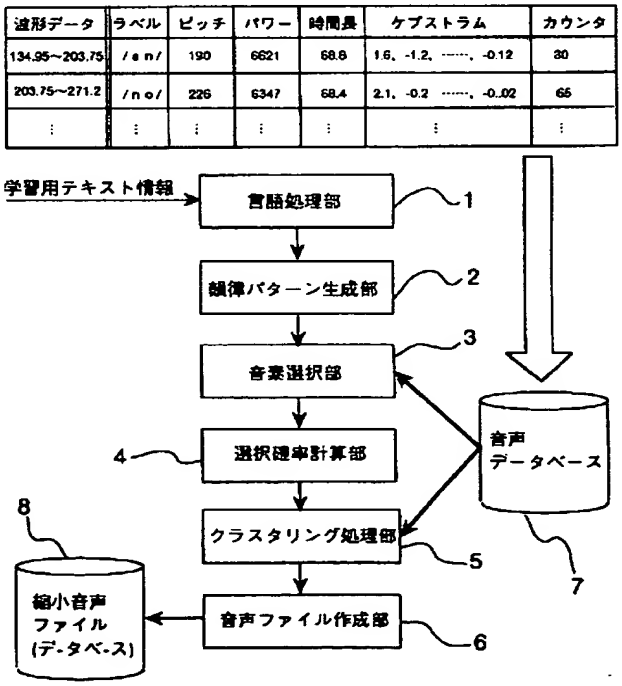
最終頁に続く

(54)【発明の名称】音声データベースにおける音素片情報最適化方法、及び音素片情報最適化装置

(57)【要約】

【課題】　従来の音声データベース最適化方法によってクラスタリング処理して音素片情報を削減したとしても、コンテキストクラスタテーブルには音声合成に際して全く使用されない音素片情報を多く含んだままの状態であるといった問題があった。

【解決手段】　本発明は、文章発話から切り出した音素片を接続することにより合成音を得る波形合成に適用される音声データベースにおける音素片情報最適化方法において、予め学習用テキスト情報を用いて合成し、その合成結果に従って前記音声データベースを構成する各音素片情報の使用頻度を求め、該使用頻度に基づいてクラスタリング処理を行うことにより、音声ファイルの音素片情報を最適化する。



1

【特許請求の範囲】

【請求項１】　文章発話から切り出した音素片を接続することにより合成音を得る波形合成に適用される音声データベースにおける音素片情報最適化方法において、
予め学習用テキスト情報を用いて合成し、その合成結果に従って前記音声データベースを構成する各音素片情報の使用頻度を求め、該使用頻度に基づいてクラスタリング処理を行うことにより、音声ファイルの音素片情報を最適化することを特徴とする音声データベースにおける音素片情報最適化方法。

【請求項２】　文章発話から切り出した音素片を接続することにより合成音を得る波形合成に適用される音声データベースにおける音素片情報最適化方法において、
音素選択部が、学習用テキスト情報を入力として、前記文章発話から切り出した音素片を蓄積した音声データベースから最適な音素片を選択する第１ステップと、
選択確率計算部が、前記音素選択部によって選択された各音素片の選択確率を求める第２ステップと、
クラスタリング処理部が、前記音声データベースに対し、所定のパラメータ空間において、前記選択確率を音素片の分布確率としてクラスタリング処理を行う第３ステップと、及び音素波形素片登録部が、前記クラスタリング処理部によってクラスタリングされた、各クラスタの中から代表音素片を選択する第４ステップ、からなることを特徴とする音声データベースにおける音素片情報最適化方法。

【請求項３】　文章発話から切り出した音素片を接続することにより合成音を得る波形合成に適用される音声データベースにおける音素片情報最適化装置において、
前記文章発話から切り出した音素片を蓄積した音声データベースと、
学習用テキスト情報を入力として、前記音声データベースからなる最適な音素片を選択する音素選択部と、
該音素選択部によって選択された、各音素片の選択確率を求める選択確率計算部と、
前記音声データベースに対し、所定のパラメータ空間において、前記選択確率を音素片の分布確率としてクラスタリング処理を行うクラスタリング処理部と、該クラスタリング処理部によってクラスタリング処理された、各クラスタの中から代表音素片を選択する音素波形素片登録部と、を備えることを特徴とする音声データベースにおける音素片情報最適化装置。

【請求項４】　前記クラスタリング処理部は、各クラスタ内のセントロイドから前記クラスタ内に含まれる全ての音素片までの距離が最小になるようにクラスタリング処理することを特徴とする請求項１、又は２記載の音声データベースにおける音素片情報最適化方法。

【請求項５】　前記学習用テキスト情報は、文章から構成されていることを特徴とする請求項１、又は２記載の音声データベースにおける音素片情報最適化方法。

2

【請求項６】　前記音素片は、少なくとも波形情報から構成されていることを特徴とする請求項１、又は２記載の音声データベースにおける音素片情報最適化方法。

【請求項７】　前記クラスタリング処理部は、各クラスタ内のセントロイドから前記クラスタ内に含まれる全ての音素片までの距離が最小になるようにクラスタリング処理することを特徴とする請求項３記載の音声データベースにおける音素片情報最適化装置。

【請求項８】　前記学習用テキスト情報は、文章から構成されていることを特徴とする請求項３記載の音声データベースにおける音素片情報最適化装置。

【請求項９】　前記音素片は、少なくとも波形情報から構成されていることを特徴とする請求項３記載の音声データベースにおける音素片情報最適化装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、予め文章発話から切り出して蓄積した、音素片情報からなる音声データベースから最適な音素片情報を選択し接続することにより合成音を得る波形合成に適用される、音声データベースにおける音素片情報最適化方法、及び音素片情報最適化装置に関する。

【０００２】

【従来の技術】従来、音声波形を接続して合成音を得る波形合成に適用される音声データベースの音素片情報に対してクラスタリングを行い、最適な音素片情報を音声ファイルに登録する音声ファイル構成方式等が特開平８－２６３５２０号公報に開示されている。

【０００３】図５は、従来のコンテキストクラスタリングの処理を示すフローチャートである。同図において、音声データベース１００内の音素ラベリングされた波形データ中から同一の音素ラベルが付与されている波形データを全て取り出し、初期クラスタ１１０とする（ステップ２０１）。

【０００４】次に、この初期クラスタ１１０内の個々の波形データ（要素）を特徴分析する（ステップ２０２）。この特徴分析においては、ＬＰＣ（線形予測符号化法）ケプストラム等の特徴パラメータの次数をｎとし、かつ、分析窓関数のフレーム周期を可変として、フレーム数がｍフレームとなるように分析を行うことにより、各要素に対してｎ×ｍ次元の特徴パラメータ行列を得る。

【０００５】次にこの特徴分析の結果を用いて、初期クラスタ１１０のクラスタ歪を求める（ステップ２０３）。具体的には、特徴パラメータのベクトル空間において、初期クラスタ１１０の全ての要素と予め求めておいたセントロイドとの間の距離の２乗和を求めて、これを初期クラスタ１１０のクラスタ歪と定義する。

【０００６】こうして初期クラスタ１１０のクラスタ歪を求め、これをコンテキストクラスタテーブル２０８に

登録する。このコンテキストクラスタテーブル２０８には、図示のように、各クラスタ毎に、それに属するコンテキストと、そのセントロイドと、そのクラスタ歪と、それに含まれる要素波形の集合とが登録されている。

【０００７】尚、初期クラスタ１１０のクラスタ歪を求めた段階では、初期クラスタ１００だけがコンテキストクラスタテーブル２０８に登録されていることになる。

【０００８】次にコンテキストクラスタテーブル２０８中からクラスタ歪が最大となるクラスタを求め（ステップ２０４）、この求めたクラスタを、コンテキストクラスタテーブル２０８中から取り出し、コンテキストにより更に２つのクラスタに分割する（ステップ２０５）。

【０００９】尚、最初の段階では、初期クラスタ１１０だけがコンテキストクラスタテーブル２０８に登録されているので、この初期クラスタ１１０に対してクラスタ分割が行われる。

【００１０】このようにして、初期クラスタ１１０の分割が行われた後、コンテキストクラスタテーブル２０８において、初期クラスタ１１０が削除され、分割された２つのクラスタが新たに登録される（ステップ２０６）。

【００１１】以上の処理（ステップ２０３～２０６）を繰り返すことにより、初期クラスタ１１０は次第に小さいクラスタに細分化されていく。そして、この各繰り返しループ毎に、コンテキストクラスタリングの終了判定が行われる（ステップ２０７）。

【００１２】

【発明が解決しようとする課題】然し乍ら、この音声データベース１００の音素片情報を削減して音声ファイル（データベース）を作成したとしても、音声データベース１００に含まれる文章と音声合成器に入力する文章とでは音素片の出現頻度が異なるため、コンテキストクラスタテーブル２０８には音声合成に際して全く使用されない音素片情報を多く含んだままの状態であるといった問題が依然残っていた。

【００１３】従って、本発明は、大量の学習用テキスト情報（文章）を予め用意し、それを全ての音素片を用いた音声合成器で予め合成し、その結果から各音素片の使用された回数（頻度情報）を求め、その分布にしたがって距離の総和を計算し、クラスタリングを行うことを特徴とする。

【００１４】これによって、クラスタリング処理の対象となっている音声データベースに様々な音素片情報が含まれていたとしても、頻繁に使用される音声に対して多くの音素片を割り当てた音声ファイル（データベース）を構築することが可能となる。

【００１５】

【課題を解決するための手段】本発明の音声データベースにおける音素片情報最適化方法は、文章発話から切り出した音素片を接続することにより合成音を得る波形合成に適用される音声データベースにおける音素片情報最適化方法において、予め学習用テキスト情報を用いて合成し、その合成結果に従って前記音声データベースを構成する各音素片情報の使用頻度を求め、該使用頻度に基づいてクラスタリング処理を行うことにより、音声ファイルの音素片情報を最適化することを特徴とする。

【００１６】また、本発明の音声データベースにおける音素片情報最適化方法は、文章発話から切り出した音素片を接続することにより合成音を得る波形合成に適用される音声データベースにおける音素片情報最適化方法において、音素選択部が、学習用テキスト情報を入力として、前記文章発話から切り出した音素片を蓄積した音声データベースから最適な音素片を選択する第１ステップと、選択確率計算部が、前記音素選択部によって選択された各音素片の選択確率を求める第２ステップと、クラスタリング処理部が、前記音声データベースに対し、所定のパラメータ空間において、前記選択確率を音素片の分布確率としてクラスタリング処理を行う第３ステップと、及び音素波形素片登録部が、前記クラスタリング処理部によってクラスタリングされた、各クラスタの中から代表音素片を選択する第４ステップ、からなることを特徴とする。

【００１７】本発明の音声データベースにおける音素片情報最適化装置は、文章発話から切り出した音素片を接続することにより合成音を得る波形合成に適用される音声データベースにおける音素片情報最適化装置において、前記文章発話から切り出した音素片を蓄積した音声データベースと、学習用テキスト情報を入力として、前記音声データベースからなる最適な音素片を選択する音素選択部と、該音素選択部によって選択された、各音素片の選択確率を求める選択確率計算部と、前記音声データベースに対し、所定のパラメータ空間において、前記選択確率を音素片の分布確率としてクラスタリング処理を行うクラスタリング処理部と、該クラスタリング処理部によってクラスタリング処理された、各クラスタの中から代表音素片を選択する音素波形素片登録部と、を備えることを特徴とする。

【００１８】また、前記クラスタリング処理部は、各クラスタ内のセントロイドから前記クラスタ内に含まれる全ての音素片までの距離が最小になるようにクラスタリング処理することを特徴とする。

【００１９】前記学習用テキスト情報は、文章から構成されていることを特徴とする。

【００２０】更に、前記音素片は、少なくとも波形情報から構成されていることを特徴とする。

【００２１】

【発明の実施の形態】本発明の実施の形態を図１～図４を用いて説明する。

【００２２】図１は、本発明を実現するための装置の概略構成図である。また、図２は、本発明における、音声

5

データベースにおける音素片情報最適化方法を実現するためのフローチャートである。

【００２３】以下、図１を参照し乍ら、図２の処理過程を説明する。

【００２４】ステップＳ１では、学習用テキスト情報（文章）が言語処理部１に入力されると、言語処理部１は、形態素解析、係り受け解析を行い、解析後の音素に対して音素記号、品詞、及びアクセント記号列を付与する。

【００２５】ステップＳ３では、韻律パターン生成部２は音素記号、アクセント記号列、及び形態素解析の結果より得られる入力テキストの品詞情報を用いて、音素中心付近での基本周波数F_0、パワー、音韻継続時間長を

6

推定する。

【００２６】ステップＳ５においては、数１に示す式を用いてコストを求め、音素を選択する。本ステップにおける具体的な音素片の選択は、ステップＳ３で推定された基本周波数F_0、パワー、音韻継続時間長の推定値との非適合を表わすコスト、及び各音素片を接続するときのコストを計算し、その総和が最小になる音素片の組み合わせをＤＰ（ダイナミックプログラミング）法に従い数１を用いることにより探索を行う。ここで、コストを示すコスト関数Ｄ（Ｆ）を数１に示す。

【００２７】

【数１】

$$D(F)=\sum_{i=1}^{k}(w_{F0}D_{F0}(c(i))+w_{pow}D_{pow}(c(i))+w_{dur}D_{dur}(c(i))+w_{posi}D_{posi}(c(i)))+\max(w^{c}_{F0}D^{c}_{F0}(c(i))+w^{c}_{pow}D^{c}_{pow}(c(i))+w^{c}_{cep}D^{c}_{cep}(c(i))+w^{c}_{ph}D^{c}_{ph}(c(i)))$$

【００２８】尚、D_{F0}、D_{pow}、D_{dur}は、音素中心付近での基本周波数、パワー、音韻継続時間長の推定値と合成単位との差であり、D_{posi}は、文中の位置（語頭、語中、及び語尾）の違いを数値化した値である。D^{c}_{F0}、D^{c}_{pow}、D^{c}_{cep}は、接続する２つの合成単位の接続点での基本周波数の差、パワーの差、ケプストラムの差である。D^{c}_{ph}は、発話環境を考慮して決定された接続の行い易さ（接続優先順位）を示す歪である。また、$w_{\cdot}$、$w^{c}_{\cdot}$は、夫々のパラメータに乗ずる重み係数である。

【００２９】次に、ステップＳ７では、全ての音素片を含む音声データベース７を用いて、学習用テキスト情報（文章）を合成し、各音素片の選択された回数を計算する。具体的には、音声データベース７の全ての音素片を、適当な音素単位に分割する。このとき、無声の子音を含む場合には、ＣＶ、ＶＣに分割し、有声の子音を含む場合には、ＶＣＶに分割している。尚、「Ｃ」とは、子音（Consonant）を表わし、また「Ｖ」とは、母音（Vowel）を表わす。

【００３０】次に、各音素単位に含まれる音素片の選択された回数を平均し、音素単位の選択回数とし、各音素単位の選択確率を計算する。尚、本発明では、１度も選択されなかった音素単位にも小さな確率を割り当てることとした。ステップＳ９においては、ステップＳ７で計算された選択確率に基づいて、クラスタリング処理部５は、分布が既知の場合のＬＢＧアルゴリズムを用いて、指定された個数にクラスタリング処理を行う。

【００３１】ここで、ステップＳ９を図３を用いて詳細に説明する。

【００３２】まず、ステップＳ２１では、指定された個数の初期代表ベクトルA_0を任意に決定する。またインデックスｍ＝０、平均歪み$D_{-1}=-\infty$とする。

【００３３】ステップＳ２３では、音声データベースの全ての音素片を最も近い代表ベクトルA_nが属するクラスタ$P(A_n)$に分割する。この時の距離の計算は数２を用いる。

【００３４】

【数２】

$$D(F)=\sum_{i=1}^{k}(w_{F0}D_{F0}(c(i))+w_{pow}D_{pow}(c(i))+w_{dur}D_{dur}(c(i))+w_{posi}D_{posi}(c(i)))$$

【００３５】ステップＳ２５では、各クラスタに属する全ての音素片に関する重心（セントロイド）を計算し、ｍ＋１の代表ベクトルとする。このセントロイドは、音素片の音響パラメータのベクトルの各要素ごとの平均を計算することで求められるが、この平均は、各音素の選択確率を用いて計算される。

【００３６】ステップＳ２７では、代表ベクトルA_{n-1}、クラスタ$P(A_{n-1})$の時の平均歪みD_{n-1}を計算

する。歪みは、前記数２を選択確率で平均した結果である。

【００３７】ステップＳ２９は、インデックスを１増加させる。

【００３８】ステップＳ３０は、終了判定を行っている。歪みの減少率を計算し一定量ε以下ならその時のクラスタＰ(A$_n$)を出力として終了する。

【００３９】ここで、図２に戻って更に説明を続ける。

【００４０】最終的に、ステップＳ１１では、ステップＳ９で求められた各クラスタのセントロイドを計算し、それに最も近い音素を選択音素として音声ファイル作成部６が登録することによって、縮小（削減）された音声ファイル（データベース）が新たに作成される。

【００４１】次に、本発明の有効性を確かめるため、評価実験を行った。本実験では、地名の読み上げを行う合成器の生成を目的とした。学習用文章には、新郵便番号データのうち九州地方を除く全てを用いた。新郵便番号データより、市・郡名称、区町村名称、町域名称を抽出し、「ここは、Ｘ市、Ｙ区、Ｚ町、です。」という文章に変換し合成を行った。

【００４２】その結果より、「ここは、」と「です。」の部分を除き、残りの結果より各音素の選択確率を求めた。求めた選択確率を用いてクラスタリングした縮小ファイル（データベース）と、選択確率が一様としてクラスタリングした縮小ファイル（データベース）を用いて、学習に用いた地名、学習に用いなかった地名(九州地方）、小説の3種類の文章を合成し評価した。

【００４３】以下に実験に用いた音声ファイル（データベース）のサイズ、および実験結果を示す。

【００４４】

【表１】

学習用テキスト	３０５７４５単語、２６３５７６５音素
音声データベース	４８０６９音素
縮小データベース	５５００音素

【００４５】図４において、縦軸は、地名20文章、小説5文章を合成した時の数１の歪コストの合計を文章の総音素数で割った1音素当りの平均歪である。また、図４中の斜線は選択確率が一様として作成した音声ファイル（データベース）による合成結果(conventional）、また交差線は提案方式による結果(proposed）、更に縦線は全ての音素片を含む音声ファイル（データベース）による合成結果(all)である。

【００４６】place-name (closed)は学習に用いた地名、place-name (open)は学習に用いなかった地名(九州地方）、novelは全く環境の異なる文章である小説の結果を示す。

【００４７】この結果より、全ての場合で提案方式の方が選択確率を一様とした場合と比較して歪が少なくなっており、提案方式が有効であることがわかる。それぞれの文章の種類ごとに比較すると、proposedの歪は地名読み上げではallに近いが、小説読み上げではconventionalに近い。これは、open-closedに関わらず言えることで、地名読み上げという環境への最適化が行われていることがわかる。

【００４８】

【発明の効果】以上の説明から明らかなように、本発明によれば、文章発話から切り出した音素片を接続することにより合成音を得る波形合成に適用される音声データベースにおける音素片情報最適化方法において、予め学習用テキスト情報を用いて合成し、その合成結果に従って前記音声データベースを構成する各音素片情報の使用頻度を求め、該使用頻度に基づいてクラスタリング処理を行うことにより、使用頻度の高い音素片情報からなる音声ファイル（データベース）を作成することができる効果を奏する。

【００４９】更に、本発明は、文章発話から切り出した音素片を接続することにより合成音を得る波形合成に適用される音声データベースにおける音素片情報最適化装置において、前記文章発話から切り出した音素片を蓄積した音声データベースと、学習用テキストを入力として、前記音声データベースからなる最適な音素片を選択する音素選択部と、該音素選択部によって選択された、各音素片の選択確率を求める選択確率計算部と、前記音声データベースに対し、所定のパラメータ空間において、前記選択確率を音素片の分布確率としてクラスタリング処理を行うクラスタリング部と、該クラスタリング部によってクラスタリングされた、各クラスタの中から代表音素片を選択する音素波形素片登録部と、を備えることにより、使用頻度の高い音声には多くの音素片情報を割り当てることが出来る効果を奏する。

【図面の簡単な説明】

【図１】本発明を実現するための装置の概略構成図である。

【図２】本発明における、音声データベースにおける音素片情報最適化方法を実現するためのフローチャートである。

【図３】図２に示すステップＳ９の処理を詳細に表したフローチャートである。

【図４】本発明の評価実験の結果を示す図である。

【図５】従来のコンテキストクラスタリングの処理を示

9　　10

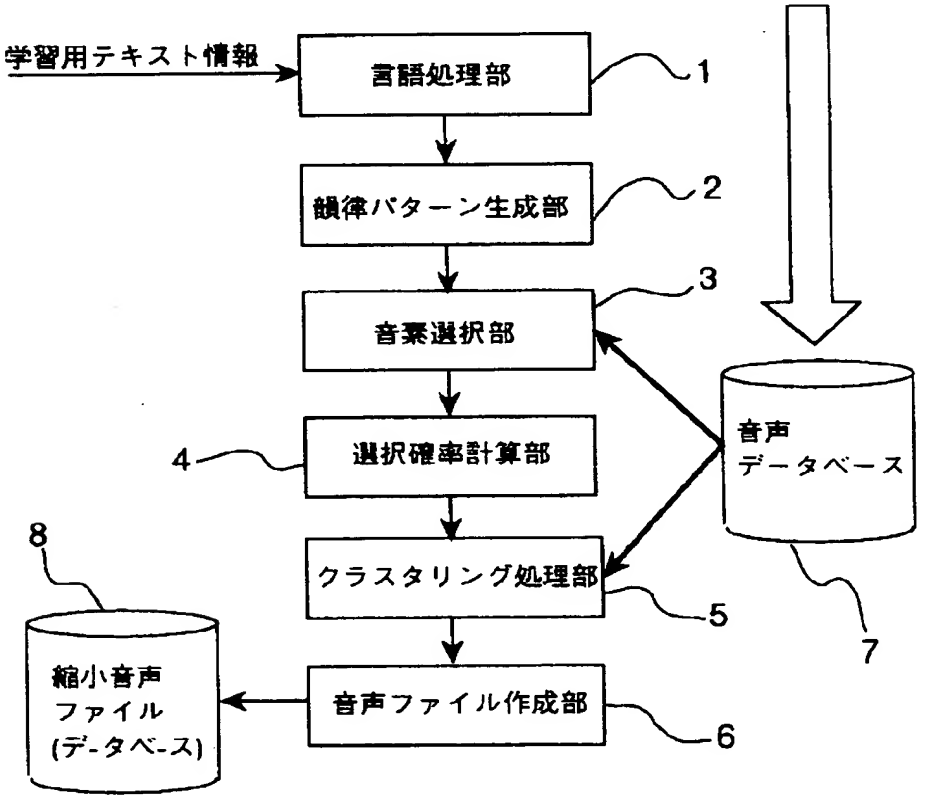
すフローチャートである。

【符号の説明】

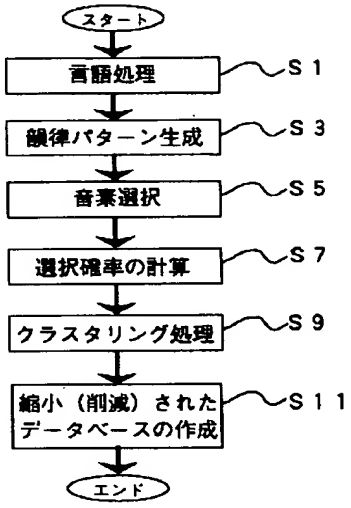
1・・・・・言語処理部
2・・・・・韻律パターン生成部
3・・・・・音素選択部
4・・・・・選択確率計算部
5・・・・・クラスタリング処理部
6・・・・・音声ファイル作成部
7・・・・・音声データベース
8・・・・・縮小音声ファイル（データベース）

【図１】

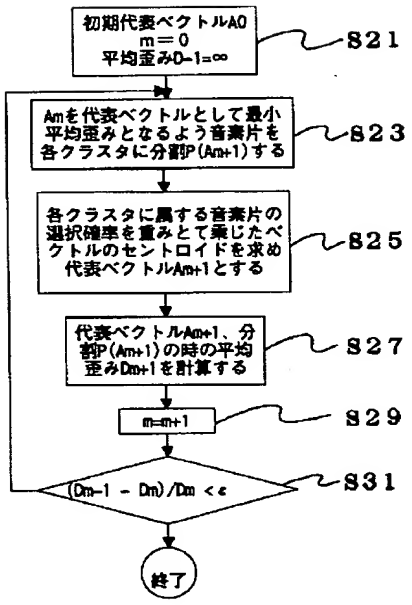
波形データ	ラベル	ピッチ	パワー	時間長	ケプストラム	カウンタ
134.95～203.75	/a n/	190	6621	68.8	1.6、-1.2、……、-0.12	30
203.75～271.2	/n o/	226	6347	68.4	2.1、-0.2 ……、-0..02	65
⋮	⋮	⋮	⋮	⋮	⋮	⋮



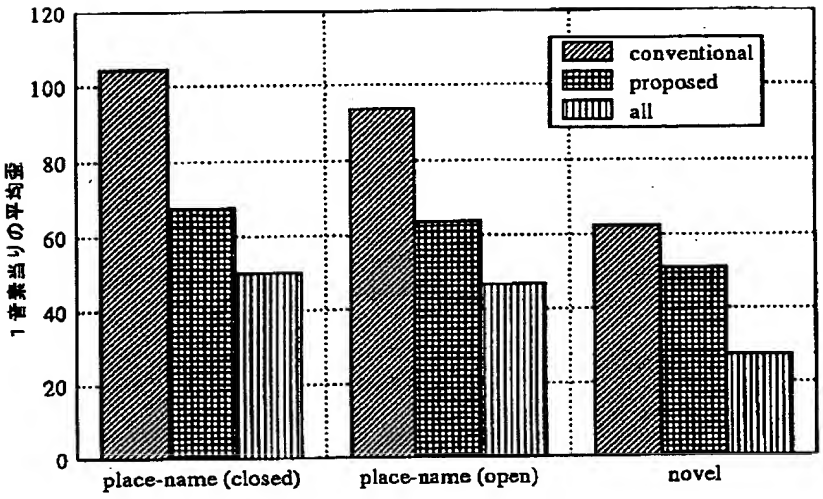
【図２】



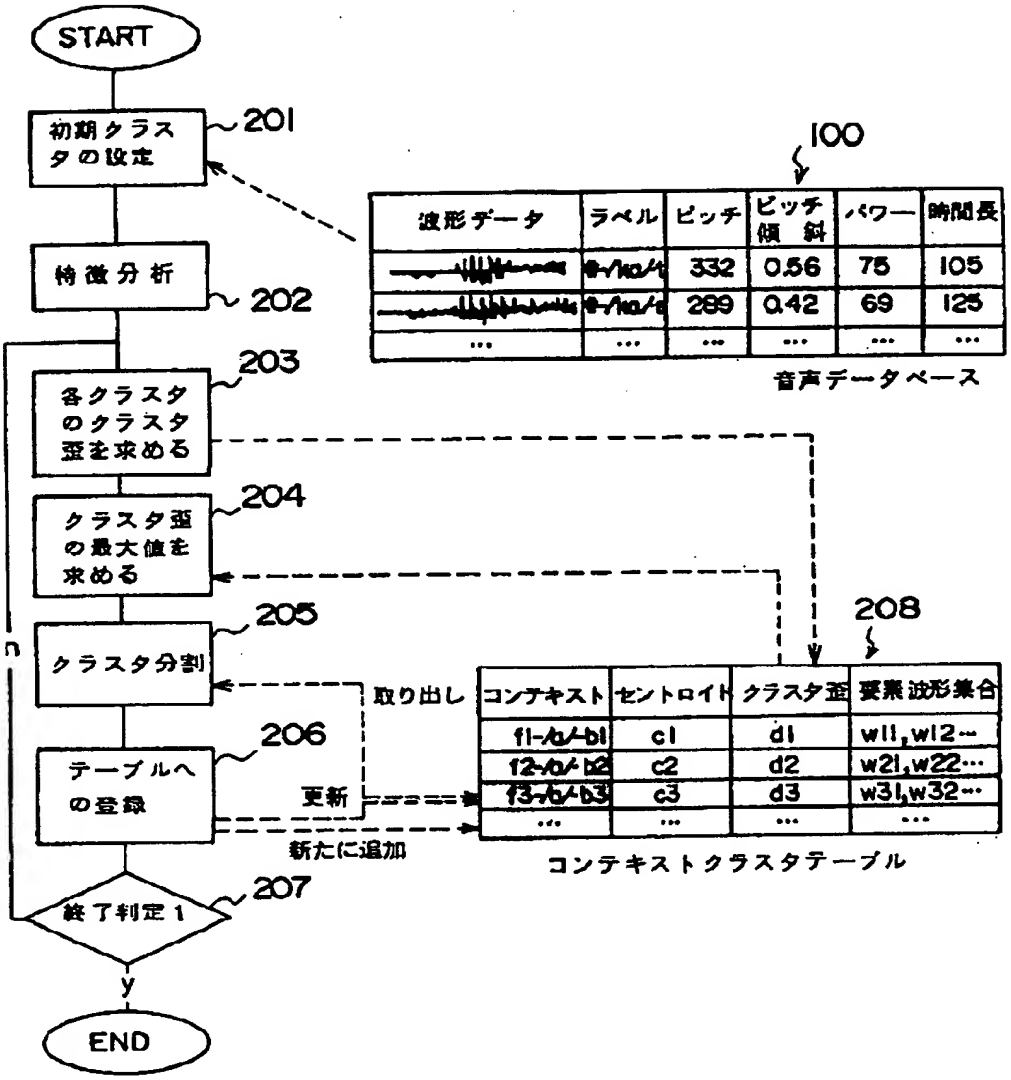
【図３】


初期代表ベクトルA0
m＝0
平均歪みD-1=∞
S21
Amを代表ベクトルとして最小平均歪みとなるよう音素片を各クラスタに分割P(Am+1)する
S23
各クラスタに属する音素片の選択確率を重みとて乗じたベクトルのセントロイドを求め代表ベクトルAm+1とする
S25
代表ベクトルAm+1、分割P(Am+1)の時の平均歪みDm+1を計算する
S27
m=m+1
S29
(Dm-1 - Dm)/Dm <ε
S31
終了

【図４】


120
100
80
60
40
20
0
1音素当りの平均歪
conventional
proposed
all
place-name (closed)
place-name (open)
novel

【図５】



フロントページの続き

(72)発明者　橋本　誠
大阪府守口市京阪本通２丁目５番５号　三洋電機株式会社内